

DELETED SEQUENCE IN M. TUBERCULOSIS, METHOD FOR DETECTING MYCOBACTERIA USING THESE SEQUENCES AND VACCINES

The present invention pertains to the field of biology, more particularly the subject of the present invention is the identification of a nucleotide sequence which make it possible in particular to distinguish an infection resulting from *Mycobacterium tuberculosis* from an infection resulting from *Mycobacterium africanum*, *Mycobacterium canetti*, *Mycobacterium microti*, *Mycobacterium bovis*, *Mycobacterium bovis BCG*. The subject of the present invention is also a method for detecting the sequences in question by the products of expression of these sequences and the kits for carrying out these methods. Finally, the subject of the present invention is novel vaccines.

Despite more than a century of research since the discovery of *Mycobacterium tuberculosis*, the aetiological agent of tuberculosis, this disease remains one of the major causes of human mortality. *M. tuberculosis* is expected to kill 3 million people annually (Snider, 1989 Rev. Inf. Dis. S335) and the number of new people getting infected each year is rising and is estimated at 8.8 million. Although the majority of these are in developing countries, the disease is assuming renewed importance in the western countries due to the increasing number of homeless people, the impact of the AIDS epidemic, the changing global migration, and the travel patterns.

Early tuberculosis often goes unrecognized in an otherwise healthy individual. Classical initial methods of diagnosis include examination of a sputum smear under a microscope for acid-fast mycobacteria and an x-ray of the lungs. However, in a vast majority of cases the sputum smear examination is negative for Mycobacteria in the early stages of the disease, and lung changes may not be obvious on an x-ray until several months following infection. Another complicating factor is that acid-fast bacteria in a sputum smear may often be other species of mycobacteria. Antibiotics used for treating tuberculosis have considerable side effects, and must be taken as a combination of three or more drugs for a six to twelve month period. In addition, the possibility of inducing the appearance of drug resistant tuberculosis prevents therapy from being administered without solid evidence to support the diagnosis. Currently the only absolutely reliable method of diagnosis is based on culturing *M. tuberculosis* from the clinical specimen and identifying it morphologically and biochemically. This usually takes anywhere from three to six weeks, during which time a patient may become seriously ill and infect other individuals. Therefore, a rapid test capable of reliably detecting the presence of *M. tuberculosis* is vital for the early detection and treatment. Several molecular tests have been developed recently for the rapid detection and

identification of *M. tuberculosis*, such as the Gen-Probe "Amplified *Mycobacterium tuberculosis* Direct Test"; this test amplifies *M. tuberculosis* 16S ribosomal RNA from respiratory specimens and uses a chemiluminescent probe to detect the amplified product with a reported sensitivity of about 91%. The discovery of the IS6110 insertion element
5 (Cave et al., Eisenach et al., 1990 J. Infectious Diseases 161:977-981; Thierry et al. 1990 J. Clin. Microbiol. 28: 2668-2673) and the belief that this element may only be present in *Mycobacterium* complex (*M. tuberculosis*, *M. bovis*, *M. bovis*-BCG, *M. africanum*, *M. canettii* and *M. microti*) spawned a whole series of rapid diagnostic strategies (Brisson-Noel et al., 1991 Lancet 338: 364-366; Clarridge et al. 1993, J. Clin. Microbiol. 31:2049-2056 ;
10 Cormican et al. 1992 J. Clin. Pathology 1992, 45 : 601-604 ; Cousins et al., 1992 J. Clin. Microbiol. 30 : 255-258 ; Del Portillo et al. 1991 J. Clin. Microbiol. 29 : 2163-2168 ; Folgueira et al., 1994 Neurology 44 :1336-1338 ; Forbes et al. 1993, J.Clin.Microbiol. 31 :1688-1694 ; Hermans et al. 1990 J. Clin. Microbiol. 28 :1204-1213 ; Kaltwasser et al. 1993 Mol. Cell. Probes 7 : 465-470 ; Kocagoz et al. 1993 J. Clin. Microbiol. 31 :1435-1438 ;
15 Kolk et al. 1992 J.Clin.Microbiol. 30 : 2567-2575 ; Kox et al. 1994 J.Clin.Microbiol. 32 :672-678 ; Liu et al. 1994 Neurology 44 :1161-1164 ; Miller et al. 1994 J. Clin.Microbiol. 32 : 393-397 ; Reischl et al. 1994 Biotechniques 17 :844-845 ; Schluger et al. 1994 Chest 105 :1116-1121 ; Shawar et al. 1993 J. Clin. Microbiol. 31: 61-65; Wilson et al 1993 J.Clin.Microbiol. 28: 2668-2673). These tests employ various techniques to extract DNA
20 from the sputum. PCR is used to amplify IS6110 DNA sequences from the extracted DNA. The successful amplification of this DNA is considered to be an indicator of the presence of *M. tuberculosis* infection. U.S. Pat. Nos. 5,168,039 and 5,370,998 have been issued to Crawford et al. for the IS6110 based detection of tuberculosis. European patent EP 0,461,045 has been issued to Guesdon for the IS6110 based detection of tuberculosis.

25 Thus, these molecular assays used to detect *M. tuberculosis* depend on the IS6110 insertion sequence (about 10 copies) or the 16S ribosomal RNA (thousands of copies). However, these methods do not provide any information regarding the sub-type of the mycobacteria. Indeed several dozen species of Mycobacteria are known, and most are non-pathogenic for humans; tuberculosis is usually caused by infection due to *M. tuberculosis*,
30 with a few cases being caused by *M. bovis*, *M. canettii*, and *M. africanum*. In order to choose an appropriate treatment and to conduct epidemiological investigations it is absolutely necessary to be able to rapidly and accurately identify isolates, i.e to distinguish the sub-type of mycobacteria of the *Mycobacterium* complex, originating from potential tuberculosis patients. That's the problem the present invention intends to solve.

The present invention provides an isolated or purified nucleic acid from *Mycobacterium* complex wherein said nucleic acid is selected from the group consisting of:

- a) SEQ ID N°1, named TbD1 region ;
- b) Nucleic acid having a sequence fully complementary to SEQ ID N°1.
- 5 c) Nucleic acid fragment comprising at least 8, 12, 15, 20, 25, 30, 50, 100, 250, 500, 750, 1000, 1500, 2000, 2500, 3000 consecutive nucleotides of SEQ ID N°1;
- d) Nucleic acid having at least 90% sequence identity after optimal alignment with a sequence defined in a) or b);
- 10 e) Nucleic acid that hybridizes under stringent conditions with the nucleic acid defined in a) or b);

As used herein, the terms « isolated » and « purified » according to the invention refer to a level of purity that is achievable using current technology. The molecules of the invention do not need to be absolutely pure (i.e., contain absolutely no molecules of other cellular macromolecules), but should be sufficiently pure so that one of ordinary skill in the art would recognize that they are no longer present in the environment in which they were originally found (i.e., the cellular middle). Thus, a purified or isolated molecule according to the present invention is one that have been removed from at least one other macromolecule present in the natural environment in which it was found. More preferably, the molecules of the invention are essentially purified and/or isolated, which means that the composition in which they are present is almost completely, or even absolutely, free of other macromolecules found in the environment in which the molecules of the invention are originally found. Isolation and purification thus does not occur by addition or removal of salts, solvents, or elements of the periodic table, but must include the removal of at least some macromolecules. The nucleic acids encompassed by the invention are purified and/or isolated by any appropriate technique known to the ordinary artisan. Such techniques are widely known, commonly practiced, and well within the skill of the ordinary artisan. As used herein, the term “ nucleic acid” refers to a polynucleotide sequence such as a single or double stranded DNA sequence, RNA sequence, cDNA sequence; such a polynucleotide sequence has been isolated, purified or synthesized and may be constituted with natural or non natural nucleotides. In a preferred embodiment the DNA molecule of the invention is a double stranded DNA molecule. As used herein, the terms "nucleic acid", "oligonucleotide", "polynucleotide" have the same meaning and are used indifferently.

By the term “*Mycobacterium* complex” as used herein, it is meant the complex of mycobacteria causing tuberculosis which are *Mycobacterium tuberculosis*, *Mycobacterium*

bovis, *Mycobacterium africanum*, *Mycobacterium microti*, *Mycobacterium canettii* and the vaccine strain *Mycobacterium bovis* BCG.

The present invention encompasses not only the entire sequence SEQ ID N°1, its complement, and its double-stranded form, but any fragment of this sequence, its complement, and its double-stranded form.

In embodiments, the fragment of SEQ ID N°1 comprises at least approximately 8 nucleotides. For example, the fragment can be between approximately 8 and 30 nucleotides and can be designed as a primer for polynucleotide synthesis. In another preferred embodiment, the fragment of SEQ ID N°1 comprises between approximately 1,500 and approximately 2,500 nucleotides, and more preferably 2153 nucleotides corresponding to SEQ ID N°4 (see figure 5). As used herein, "nucleotides" is used in reference to the number of nucleotides on a single-stranded nucleic acid. However, the term also encompasses double-stranded molecules. Thus, a fragment comprising 2,153 nucleotides according to the invention is a single-stranded molecule comprising 2,153 nucleotides, and also a double stranded molecule comprising 2153 base pairs (bp).

In a preferred embodiment, the nucleic acid fragment of the invention is specifically deleted in the genome of *Mycobacterium tuberculosis*, excepted in *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene katG and having no or very few IS6110 sequences inserted in their genome and present in the genome of *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*, *Mycobacterium bovis*, *Mycobacterium bovis* BCG. By the term "few IS6110 sequences inserted in the genome", it is meant less than ten copies in the genome of *M. tuberculosis*, more preferably less than 5 copies, for example less than two copies.

The nucleic acid fragment of the invention is preferably selected from the group consisting of:

- a) SEQ ID N°4;
- b) Nucleic acid having a sequence fully complementary to SEQ ID N°4.
- c) Nucleic acid fragment comprising at least 8, 12, 15, 20, 25, 30, 50, 100, 250, 500, 750, 1000, 1500, 2000, 2500, 3000 consecutive nucleotides of SEQ ID N°4;
- d) Nucleic acid having at least 90% sequence identity after optimal alignment with a sequence defined in a) or b);
- e) Nucleic acid that hybridizes under stringent conditions with the nucleic acid defined in a) or b).

In embodiments, the stringent conditions under which a sequence according to the invention is determined are conditions which are no less stringent than 5X SSPE, 2X

Denhardt's solution, and 0.5% (w/v) sodium dodecyl sulfate at 65°C. More stringent conditions can be utilized by the ordinary artisan, and the proper conditions for a given assay can be easily and rapidly determined without undue or excessive experimentation. As an illustrative embodiment, the stringent hybridization conditions used in order to specifically
5 detect a polynucleotide according to the present invention are advantageously the following: pre-hybridization and hybridization are performed at 65°C in a mixture containing:

- 5X SSPE (1X SSPE is 3 M NaCl, 30 mM tri-sodium citrate)
- 2X Denhardt's solution
- 0.5% (w/v) sodium dodecyl sulfate (SDS)
- 10 - 100 µg ml⁻¹ salmon sperm DNA.

The washings are performed as follows:

- two washings at laboratory temperature (approximately 21-25°C) for 10 min. in the presence of 2X SSPE and 0.1% SDS; and
- one washing at 65°C for 15 min. in the presence of 1X SSPE and 0.1% SDS.

15 The invention also encompasses the isolated or purified nucleic acid of the invention wherein said nucleic acid comprises at least a deletion of a nucleic acid fragment as defined above. Preferably, such an isolated or purified nucleic acid of the invention is the SEQ ID N°21 that corresponds to SEQ ID N°1 in which SEQ ID N°4 is deleted (absent).

20 Polynucleotides of the invention can be characterized by the percentage of identity they show with the sequences disclosed herein. For example, polynucleotides having at least 90% identity with the polynucleotides of the invention, particularly those sequences of the sequence listing, are encompassed by the invention. Preferably, the sequences show at least
25 90% identity with those of the sequence listing. More preferably, they show at least 92% identity, for example 95% or 99% identity. The skilled artisan can identify sequences according to the invention through the use of the sequence analysis software BLAST (see for example, Coffin et al., eds., "*Retroviruses*", Cold Spring Harbor Laboratory Press, pp. 723-755). Percent identity is calculated using the BLAST sequence analysis program suite,
30 Version 2, available at the NCBI (NIH). All default parameters are used. BLAST (Basic Local Alignment Search Tool) is the heuristic search algorithm employed by the programs blastp, blastn, blastx, tblastn and tblastx, all of which are available through the BLAST analysis software suite at the NCBI. These programs ascribe significance to their findings using the statistical methods of Karlin and Altschul (1990, 1993) with a few enhancements.

Using this publicly available sequence analysis program suite, the skilled artisan can easily identify polynucleotides according to the present invention.

It is well within the skill of the ordinary artisan to identify regions of the nucleic acid sequence of the invention, which would be useful as a probe, primer, or other experimental, diagnostic, or therapeutic aid. For example, the ordinary artisan could utilize any of the widely available sequence analysis programs to select regions (fragments) of these sequences that are useful for hybridization assays such as Southern blots, Northern blots, DNA binding assays, and/or *in vitro*, *in situ*, or *in vivo* hybridizations. Additionally, the ordinary artisan, with the sequences of the present invention, can utilize widely available sequence analysis programs to identify regions that can be used as probes and primers, as well as for design of anti-sense molecules. The only practical limitation on the fragment chosen by the ordinary artisan is the ability of the fragment to be useful for the purpose for which it is chosen. For example, if the ordinary artisan wished to choose a hybridization probe, he would know how to choose one of sufficient length, and of sufficient stability, to give meaningful results. The conditions chosen would be those typically used in hybridization assays developed for nucleic acid fragments of the approximate chosen length.

Thus, the present invention provides short oligonucleotides, such as those useful as probes and primers. In embodiments, the probe and/or primer comprises 8 to 30 consecutive nucleotides of the polynucleotide according to the invention or the polynucleotide complementary thereto. Advantageously, a fragment as defined herein has a length of at least 8 nucleotides, which is approximately the minimal length that has been determined to allow specific hybridization. Preferably the nucleic fragment has a length of at least 12 nucleotides and more preferably 20 consecutive nucleotides of any of SEQ ID N°1 or SEQ ID N°4. The sequence of the oligonucleotide can be any of the many possible sequences according to the invention. Preferably, the sequence is selected from the following group SEQ ID N° 13, SEQ ID N° 14, SEQ ID N°15, SEQ ID N°16, SEQ ID N°17, SEQ ID N°18. More precisely, the primers SEQ ID N°13, SEQ ID N°14, SEQ ID N°15 and SEQ ID N°16 are contained in the nucleic acid fragment SEQ ID N°4. The primers SEQ ID N°17 and SEQ ID N°18 are contained in the nucleic acid sequence SEQ ID N°1 and are flanking the nucleic acid fragment of SEQ ID N°4 (see figure 5).

Thus, the polynucleotides of SEQ ID N°1 and SEQ ID N°4, and their fragments, can be used to select nucleotide primers, notably for an amplification reaction, such as the amplification reactions further described.

PCR is described in US Patent No. 4,683,202, which is incorporated in its entirety herein. The amplified fragments may be identified by agarose or polyacrylamide gel

electrophoresis, by a capillary electrophoresis, or alternatively by a chromatography technique (gel filtration, hydrophobic chromatography, or ion exchange chromatography). The specificity of the amplification can be ensured by a molecular hybridization using as nucleic probes the polynucleotides of SEQ ID N°1 or SEQ ID N°4, and their fragments, oligonucleotides that are complementary to these polynucleotides or fragments thereof, or their amplification products themselves, and/or even by DNA sequencing.

The following other techniques related to nucleic acid amplification may also be used and are generally preferred to the PCR technique. The Strand Displacement Amplification (SDA) technique is an isothermal amplification technique based on the ability of a restriction enzyme to cleave one of the strands at a recognition site (which is under a hemiphosphorothioate form) and on the property of a DNA polymerase to initiate the synthesis of a new strand from the 3'OH end generated by the restriction enzyme and on the property of this DNA polymerase to displace the previously synthesized strand being localized downstream. The SDA amplification technique is more easily performed than PCR (a single thermostatted water bath device is necessary), and is faster than the other amplification methods. Thus, the present invention also comprises using the nucleic acid fragments according to the invention (primers) in a method of DNA or RNA amplification according to the SDA technique.

When the target polynucleotide to be detected is a RNA, for example a mRNA, a reverse transcriptase enzyme will be used before the amplification reaction in order to obtain a cDNA from the RNA contained in the biological sample. The generated cDNA is subsequently used as the nucleic acid target for the primers or the probes used in an amplification process or a detection process according to the present invention.

The non-labeled polynucleotides or oligonucleotides of the invention can be directly used as probes. Nevertheless, the polynucleotides or oligonucleotides are generally labeled with a radioactive element (^{32}P , ^{35}S , ^3H , ^{125}I) or by a non-isotopic molecule (for example, biotin, acetylaminofluorene, digoxigenin, 5-bromodesoxyuridine, fluorescein) in order to generate probes that are useful for numerous applications. Examples of non-radioactive labeling of nucleic acid fragments are described in French patent N° FR 78 10975 and by Urdea *et al.* (1988, *Nucleic Acids Research* 11:4937-4957) or Sanchez-Pescador *et al.* (1988, *J. Clin. Microbiol.* 26(10):1934-1938), the disclosures of which are hereby incorporated in their entirety. Other labeling techniques can also be used, such as those described in French patents FR 2 422 956 and FR 2 518 755. The hybridization step may be performed in different ways. See, for example, Matthews *et al.*, 1988, *Anal. Biochem.* 169:1-25. A general method comprises immobilizing the nucleic acid that has been extracted from the biological

sample on a substrate (for example, nitrocellulose, nylon, polystyrene) and then incubating, in defined conditions, the target nucleic acid with the probe. Subsequent to the hybridization step, the excess amount of the specific probe is discarded and the hybrid molecules formed are detected by an appropriate method (radioactivity, fluorescence or enzyme activity measurement, etc.).

Amplified nucleotide fragments are useful, among other things, as probes used in hybridization reactions in order to detect the presence of one polynucleotide according to the present invention or in order to detect mutations. The primers may also be used as oligonucleotide probes to specifically detect a polynucleotide according to the invention.

The oligonucleotide probes according to the present invention may also be used in a detection device comprising a matrix library of probes immobilized on a substrate, the sequence of each probe of a given length being localized in a shift of one or several bases, one from the other, each probe of the matrix library thus being complementary to a distinct sequence of the target nucleic acid. Optionally, the substrate of the matrix may be a material able to act as an electron donor, the detection of the matrix positions in which an hybridization has occurred being subsequently determined by an electronic device. Such matrix libraries of probes and methods of specific detection of a target nucleic acid is described in the European patent application N° EP-0 713 016 (Affymax technologies) and also in the US patent N° US-5,202,231 (Drmanac). Since almost the whole length of a mycobacterial chromosome is covered by BAC-based genomic DNA library (i.e. 97% of the *M. tuberculosis* chromosome is covered by the BAC library I-1945), these DNA libraries will play an important role in a plurality of post-genomic applications, such as in mycobacterial gene expression studies where the canonical set of BACs could be used as a matrix for hybridization studies. Thus it is also in the scope of the invention to provide a nucleic acid chips, more precisely a DNA chips or a protein chips that respectively comprises a nucleic acid or a polypeptide of the invention.

The present invention is also providing a vector comprising the isolated DNA molecule of the invention. A "vector" is a replicon in which another polynucleotide segment is attached, so as to bring the replication and/or expression to the attached segment. A vector can have one or more restriction endonuclease recognition sites at which the DNA sequences can be cut in a determinable fashion without loss of an essential biological function of the vector, and into which a DNA fragment can be spliced in order to bring about its replication and cloning. Vectors can further provide primer sites (e.g. for PCR), transcriptional and/or translational initiation and/or regulation sites, recombinational signals, replicons, selectable markers, etc. Beside the use of homologous recombination or restriction enzymes to insert a

desired DNA fragment into the vector, UDG cloning of PCR fragments (US Pat. No. 5,334,575), T:A cloning, and the like can also be applied. The cloning vector can further contain a selectable marker suitable for use in the identification of cells transformed with the cloning vector.

5 The vector can be any useful vector known to the ordinary artisan, including, but not limited to, a cloning vector, an insertion vector, or an expression vector. Examples of vectors include plasmids, phages, cosmids, phagemid, yeast artificial chromosome (YAC), bacterial artificial chromosome (BAC), human artificial chromosome (HAC), viral vector, such as adenoviral vector, retroviral vector, and other DNA sequences which are able to replicate or
10 to be replicated *in vitro* or in a host cell, or to convey a desired DNA segment to a desired location within a host cell.

According to a preferred embodiment of the invention, the recombinant vector is a BAC pBeloBAC11 in which the genomic region of *Mycobacterium bovis-BCG* 1173P3 that spans the region corresponding to the locus 1,760,753 bp to 1,830,364 bp in the genome of *M.*
15 *tuberculosis* H37Rv has been inserted into the HindIII restriction site; this recombinant vector is named X229. In this region, the inventors have demonstrated the deletion of a 2153 bp fragment, corresponding to SEQ ID N°4, in the vast majority of *M. tuberculosis* strains excepted strains of *M. tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome. That's the reason why the
20 inventors named this deletion of 2153 bp TbD1 ("*M. tuberculosis* specific deletion 1"). TbD1 is flanked by the sequence GGC CTG GTC AAA CGC GGC TGG ATG CTG and AGA TCC GTC TTT GAC ACG ATC GAC G. External primers hybridizing with such sequences outside TbD1 or the complementary sequences thereof can be used for the amplification of TbD1 to check for the presence or the absence of the deletion of the TbD1.
25 The inventors design for example the following primers:

5'- CTA CCT CAT CTT CCG GTC CA-3' (SEQ ID N°17)

5'- CAT AGA TCC CGG ACA TGG TG-3' (SEQ ID N°18)

In order to get a specific 500 pb probe for hybridization experiments, a PCR amplification of a fragment comprised in TbD1 may be realized by using the plasmid X229 as a matrix. The
30 amplification of a fragment of approximatively 500 bp contained in TbD1 can be performed by using the following primers:

5'- CGT TCA ACC CCA AAC AGG TA-3' (SEQ ID N°13)

5'- AAT CGA ACT CGT GGA ACA CC-3' (SEQ ID N°14)

The amplification of a fragment of approximatively 2,000 bp contained in TbD1 can be
35 performed by using the following primers:

5'- ATT CAG CGT CTA TCG GTT GC-3' (SEQ ID N°15)

5'- AGC AGC TCG GGA TAT CGT AG-3' (SEQ ID N°16)

The PCR conditions are the following: denaturation 95°C 1 min, then 35 cycles of amplification [95°C during 30 seconds, 58°C during 1 min] , then elongation 72°C during 4 min.

Thus, this invention also concerns a recombinant cell host which contains a polynucleotide or recombinant vector according to the invention. The cell host can be transformed or transfected with a polynucleotide or recombinant vector to provide transient, stable, or controlled expression of the desired polynucleotide. For example, the polynucleotide of interest can be subcloned into an expression plasmid at a cloning site downstream from a promoter in the plasmid and the plasmid can be introduced into a host cell where expression can occur. The recombinant host cell can be any suitable host known to the skilled artisan, such as a eukaryotic cell or a microorganism. For example, the host can be a cell selected from the group consisting of *Escherichia coli*, *Bacillus subtilis*, insect cells, and yeasts. According to a preferred embodiment of the invention, the recombinant cell host is a commercially available *Escherichia coli* DH10B (Gibco) containing the BAC named X229 previously described. This *Escherichia coli* DH10B (Gibco) containing the BAC named X229 has been deposited with the Collection Nationale de Cultures de Microorganismes (CNCM), Institut Pasteur, Paris, France, on February 18th, 2002 under number CNCM I-2799.

Another aspect of the invention is the product of expression of all or part of the nucleic acid according to the invention, including the nucleic acid fragment specifically deleted in the genome of *Mycobacterium tuberculosis*, excepted in *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome as defined previously. The expression "product of expression" is understood to mean any isolated or purified protein, polypeptide or polypeptide fragment resulting from the expression of all or part of the above-mentioned nucleotide sequences. Among those product of expression, one can cite the membrane protein mmpL6 corresponding to SEQ ID N°6, the membrane protein mmpS6 corresponding to SEQ ID N°3 or SEQ ID N°10 (the two sequences SEQ ID N°3 and SEQ ID N°10 are identical), and their truncated or rearranged forms due to the deletion of a nucleic acid fragment according to the invention. For example, SEQ ID N°8 is a truncated form of mmpL6 protein, SEQ ID N°12 is a truncated form of mmpS6 protein and SEQ ID N°22 is a fusion product [mmpS6-mmpL6] of both rearranged mmpL6 and mmpS6 proteins.

It is now easy to produce proteins in large amounts by genetic engineering techniques through the use of expression vectors, such as plasmids, phages, and phagemids. The polypeptide of the present invention can be produced by insertion of the appropriate polynucleotide into an appropriate expression vector at the appropriate position within the vector. Such manipulation of polynucleotides is well known and widely practiced by the
5 ordinary artisan. The polypeptide can be produced from these recombinant vectors either *in vitro* or *in vivo*. All the isolated or purified nucleic acids encoding the polypeptide of the invention are in the scope of the invention. The polypeptide of the invention is a polypeptide encoded by a polynucleotide which hybridizes to any of SEQ ID N°1 or N°4 under stringent
10 conditions, as defined herein.

More preferably, said isolated or purified nucleic acid according the invention is selected among:

- the *mmpL6* gene of sequence SEQ ID N°5 contained in SEQ ID N°1 and encoding the *mmpL6* protein of sequence SEQ ID N°6;
- 15 - the truncated form of *mmpL6* gene of sequence SEQ ID N°7 contained in TbD1 of sequence SEQ ID N°4 and encoding a truncated form of *mmpL6* protein of sequence SEQ ID N°8;
- the *mmpS6* gene of sequence SEQ ID N°9 contained in SEQ ID N°1 and encoding the *mmpS6* protein of SEQ ID N°10;
- 20 - the truncated form of *mmpS6* gene of sequence SEQ ID N°11 contained in TbD1 of sequence SEQ ID N°4 and encoding a truncated form of *mmpS6* protein of SEQ ID N°12.
- the chimeric gene of SEQ ID N°21 issued from fusion of both truncated *mmpS6* and *mmpL6* genes due to the deletion of TbD1 in the genome of *M. tuberculosis* excepted
25 strains of *M. tuberculosis* having the sequence CTG at codon 463 of gene *kaiG* and having no or very few IS6110 sequences inserted in their genome. This chimeric gene encodes the fusion polypeptide [*mmpS6-mmpL6*] of sequence SEQ ID N°22.

The present invention also provides a method for the discriminatory detection and
30 identification of:

- *Mycobacterium tuberculosis* excepted *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *kaiG* and having no or very few IS6110 sequences inserted in their genome; versus,
- *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*,
35 *Mycobacterium bovis*, *Mycobacterium bovis BCG* in a biological sample,

comprising the following steps:

- a) isolation of the DNA from the biological sample to be analyzed or production of a cDNA from the RNA of the biological sample,
- b) detection of the nucleic acid sequences of the mycobacterium present in said
5 biological sample,
- c) analysis for the presence or the absence of a nucleic acid fragment specifically deleted in the genome of *Mycobacterium tuberculosis*, excepted in *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, as
10 previously described.

By a biological sample according to the present invention, it is notably intended a biological fluid, such as sputum, saliva, plasma, blood, urine or sperm, or a tissue, such as a biopsy.

Analysis of the desired sequences may, for example, be carried out by agarose gel
15 electrophoresis. If the presence of a DNA fragment migrating to the expected site is observed, it can be concluded that the analyzed sample contained mycobacterial DNA. This analysis can also be carried out by the molecular hybridization technique using a nucleic probe. This probe will be advantageously labeled with a nonradioactive (cold probe) or radioactive element. Advantageously, the detection of the mycobacterial DNA sequences
20 will be carried out using nucleotide sequences complementary to said DNA sequences. By way of example, they may include labeled or nonlabeled nucleotide probes; they may also include primers for amplification. The amplification technique used may be PCR but also other alternative techniques such as the SDA (Strand Displacement Amplification) technique, the TAS technique (Transcription-based Amplification System), the NASBA
25 (Nucleic Acid Sequence Based Amplification) technique or the TMA (Transcription Mediated Amplification) technique.

The primers in accordance with the invention have a nucleotide sequence chosen from the group comprising SEQ ID N° 13, SEQ ID N° 14, SEQ ID N°15, SEQ ID N°16, SEQ ID N°17, SEQ ID N°18. The primers SEQ ID N°13, SEQ ID N°14, SEQ ID N°15 and SEQ ID
30 N°16 are contained in the nucleic acid fragment SEQ ID N°4, and the primers SEQ ID N°17 and SEQ ID N°18 are contained in the nucleic acid of the invention SEQ ID N°1 but not in the nucleic acid fragment SEQ ID N°4.

In a variant, the subject of the invention is also a method for the discriminatory detection and identification of:

- *Mycobacterium tuberculosis* excepted *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome; versus,

- *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*,
5 *Mycobacterium bovis*, *Mycobacterium bovis BCG* in a biological sample,
comprising the following steps:

a) bringing the biological sample to be analyzed into contact with at least one pair of primers as defined above, the DNA contained in the sample having been, where appropriate, made accessible to the hybridization beforehand,

10 b) amplification of the DNA of the mycobacterium,

c) visualization of the amplification of the DNA fragments.

The amplified fragments may be identified by agarose or polyacrylamide gel electrophoresis by capillary electrophoresis or by a chromatographic technique (gel filtration, hydrophobic chromatography or ion-exchange chromatography). The specification of the
15 amplification may be controlled by molecular hybridization using probes, plasmids containing these sequences or their product of amplification. The amplified nucleotide fragments may be used as reagent in hybridization reactions in order to detect the presence, in a biological sample, of a target nucleic acid having sequences complementary to those of said amplified nucleotide fragments. These probes and amplicons may be labeled or
20 otherwise with radioactive elements or with nonradioactive molecules such as enzymes or fluorescent elements.

The subject of the present invention is also a kit for the discriminatory detection and identification of:

- *Mycobacterium tuberculosis* excepted *Mycobacterium tuberculosis* strains having the
25 sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome; versus,

- *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*,
Mycobacterium bovis, *Mycobacterium bovis BCG* in a biological sample,
in a biological sample comprising the following elements:

30 a) at least one pair of primers as defined previously,

b) the reagents necessary to carry out a DNA amplification reaction,

c) optionally, the necessary components which make it possible to verify or compare the sequence and/or the size of the amplified fragment.

Indeed, in the context of the present invention, depending on the pair of primers
35 used, it is possible to obtain very different results. Thus, the use of primers which are

contained in the TbD1 deletion, such as for example SEQ ID N°13, SEQ ID N°14, SEQ ID N°15, SEQ ID N°16, is such that no amplification product is detectable in *M. tuberculosis* excepted in strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences in their genome, and that amplification product is detectable in

5 *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*, *Mycobacterium bovis*, *Mycobacterium bovis BCG*, *Mycobacterium tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome. The use of a pair of primers outside the TbD1 deletion such as SEQ ID N°17 and

10 SEQ ID N°18 is likely to give rise to an amplicon in *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*, *Mycobacterium bovis*, *Mycobacterium bovis BCG*, *Mycobacterium tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, of about 2100 bp whereas the use of the pair of primers outside the TbD1 deletion will give rise in

15 *M. tuberculosis* excepted in strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, to an amplicon of about few bp.

More generally, the invention pertains to the use of at least one pair of primers as defined previously for the amplification of a DNA sequence from *Mycobacterium tuberculosis* or *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*,

20 *Mycobacterium bovis*, *Mycobacterium bovis BCG*, *Mycobacterium tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome.

Indeed, the subject of the present invention is also a method for the *in vitro*

25 discriminatory detection of antibodies directed against *Mycobacterium tuberculosis* excepted *Mycobacterium tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome versus antibodies directed against *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*, *Mycobacterium bovis*, *Mycobacterium bovis BCG*, *Mycobacterium tuberculosis* having the

30 sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, in a biological sample, comprising the following steps:

a) bringing the biological sample into contact with at least one product of expression of all or part of the nucleic acid fragment specifically deleted in *M. tuberculosis* excepted in strains of *M. tuberculosis* having the sequence CTG at codon 463 of gene *katG*

35 and having no or very few IS6110 sequences inserted in their genome, as previously defined,

- b) detecting the antigen-antibody complex formed.

The subject of the present invention is also a method for the *in vitro* discriminatory detection of a vaccination with *Mycobacterium bovis* BCG, an infection by *M. bovis*, *M. canettii*, *M. microti*, *M. africanum* or *M. tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, versus an infection by *Mycobacterium tuberculosis*, excepted by *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome in a mammal, comprising the following steps:

- 10 a) preparation of a biological sample containing cells, more particularly cells of the immune system of said mammal and more particularly T cells,
- b) incubation of the biological sample of step a) with at least one product of expression of all or part of the nucleic acid fragment specifically deleted in *M. tuberculosis* excepted in strains of *M. tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, as previously defined,
- 15 c) detection of a cellular reaction indicating prior sensitization of the mammal to said product, in particular cell proliferation and/or synthesis of proteins such as gamma-interferon. Cell proliferation may be measured, for example, by incorporating ³H-Thymidine.

The invention also relates to a kit for the *in vitro* discriminatory diagnosis of a vaccination with *M. bovis* BCG, an infection by *M. bovis*, *M. canettii*, *M. microti*, *M. africanum* versus an infection by *M. tuberculosis* excepted by strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, in a mammal comprising:

- a) a product of expression of all or part of the nucleic acid fragment specifically deleted in *M. tuberculosis* excepted in strains of *M. tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, as previously defined ,
- 25 b) where appropriate, the reagents for the constitution of the medium suitable for the immunological reaction,
- 30 c) the reagents allowing the detection of the antigen-antibody complexes produced by the immunological reaction,
- d) where appropriate, a reference biological sample (negative control) free of antibodies recognized by said product,
- e) where appropriate, a reference biological sample (positive control) containing a predetermined quantity of antibodies recognized by said product.
- 35

The reagents allowing the detection of the antigen-antibody complexes may carry a marker or may be capable of being recognized in turn by a labeled reagent, more particularly in the case where the antibody used is not labeled.

The subject of the invention is also mono- or polyclonal antibodies, their chimeric
5 fragments or antibodies, capable of specifically recognizing a product of expression in accordance with the present invention.

The present invention therefore also relates to a method for the *in vitro* discriminatory detection of the presence of an antigen of *Mycobacterium tuberculosis* excepted of strains having the sequence CTG at codon 463 of gene *katG* and having no or
10 very few IS6110 sequences inserted in their genome, versus the presence of an antigen of *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*, *Mycobacterium bovis*, *Mycobacterium bovis*-BCG and *Mycobacterium tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, in a biological sample comprising the following steps:

- 15 a) bringing the biological sample into contact with an antibody of the invention,
 b) detecting the antigen-antibody complex formed.

The invention also relates to a kit for the discriminatory detection of the presence of an antigen of *Mycobacterium tuberculosis* excepted strains of *M. tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences
20 inserted in their genome versus the presence of an antigen of *Mycobacterium africanum*, *Mycobacterium canettii*, *Mycobacterium microti*, *Mycobacterium bovis*, *Mycobacterium bovis* BCG, *Mycobacterium tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, in a biological sample comprising the following steps:

- 25 a) an antibody as previously claimed ,
 b) the reagents for constituting the medium suitable for the immunological reaction,
 c) the reagents allowing the detection of the antigen-antibody complexes produced by the immunological reaction.

The above-mentioned reagents are well known to a person skilled in the art who will
30 have no difficulty adapting them to the context of the present invention.

The subject of the invention is also an immunogenic composition, characterized in that it comprises at least one product of expression in accordance with the invention. Such an immunogenic composition will be used to protect animals and humans against infections by *M. africanum*, *M. bovis*, *M. canettii*, *M. microti* and *M. tuberculosis*.

In a particular embodiment, such an immunogenic composition will comprise a product of expression of all or part of the nucleic fragment specifically deleted in the genome of *Mycobacterium tuberculosis*, excepted in *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome. And in a preferable embodiment, such an immunogenic composition will comprise a product of expression of all or part of TbD1. In this case, such an immunogenic composition will be used to protect animals and humans against infections by *M. africanum*, *M. bovis*, *M. canettii*, *M. microti* and *M. tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome.

In an other particular embodiment, such an immunogenic composition will comprise the fusion product [mmpS6-mmpL6] of SEQ ID N°22. This fusion product is due to the absence of TbD1 in *M. tuberculosis* excepted strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome. An immunogenic composition comprising this fusion product will be used to protect animals and humans specifically against infection by the vast majority of *M. tuberculosis* strains excepted strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome.

Advantageously, the immunogenic composition in accordance with the invention enters into the composition of a vaccine when it is provided in combination with a pharmaceutically acceptable vehicle and optionally with one or more immunity adjuvant(s) such as alum or a representative of the family of muramylpeptides or incomplete Freund's adjuvant.

The invention also relates to a vaccine comprising at least one product of expression in accordance with the invention in combination with a pharmaceutically compatible vehicle and, where appropriate, one or more appropriate immunity adjuvant(s).

The invention also provide an in vitro method for the detection and identification of *Mycobacterium tuberculosis* excepted *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome in a biological sample, comprising the following steps:

- a) isolation of the DNA from the biological sample to be analyzed or production of a cDNA from the RNA of the biological sample,
- b) detection of the nucleic acid sequences of the mycobacterium present in said biological sample,

- c) analysis for the presence or the absence of a nucleic acid fragment of the invention.

In another embodiment, the invention provides an *in vitro* method for the detection and identification of *Mycobacterium tuberculosis* excepted *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome in a biological sample, comprising the following steps:

- a) bringing the biological sample to be analyzed into contact with at least one pair of primers selected among nucleic acid fragments of the invention, and more preferably selected among the primers chosen from the group comprising SEQ ID N°13, SEQ ID N°14, SEQ ID N°15, SEQ ID N°16, SEQ ID N°17, SEQ ID N°18, the DNA contained in the sample having been, where appropriate, made accessible to the hybridization beforehand,
- b) amplification of the DNA of the mycobacterium,
- c) visualization of the amplification of the DNA fragments.

The invention also provides a kit for the detection and identification of *Mycobacterium tuberculosis* excepted *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome in a biological sample, comprising the following elements:

- a) at least one pair of primers selected among nucleic acid fragments of the invention, and more preferably selected among the primers chosen from the group comprising SEQ ID N°13, SEQ ID N°14, SEQ ID N°15, SEQ ID N°16, SEQ ID N°17, SEQ ID N°18,
- b) the reagents necessary to carry out a DNA amplification reaction,
- c) optionally, the necessary components which make it possible to verify or compare the sequence and/or the size of the amplified fragment.

The invention also relates to a method for the *in vitro* detection of antibodies directed against *Mycobacterium tuberculosis* excepted *Mycobacterium tuberculosis* strains having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, in a biological sample, comprising the following steps:

- a) bringing the biological sample into contact with at least one product of expression of all or part of the nucleic acid fragment specifically deleted in *M. tuberculosis* excepted in strains of *M. tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome,
- b) detecting the antigen-antibody complex formed.

It is also a goal of the invention to use the TbD1 deletion as a genetic marker for the differentiation of *Mycobacterium* strains of *Mycobacterium* complex.

It is also a goal of the invention to use mmpL6⁵⁵¹ polymorphism as a genetic marker for the differentiation of *Mycobacterium* strains of *Mycobacterium* complex.

5 The use of such genetic marker(s) in association with at least one genetic marker selected among RD1, RD2, RD3, RD4, RD5, RD6, RD7, RD8, RD9, RD10, RD11, RD13, RD14, RvD1, RvD2, RvD3, RvD4, RvD5, katG⁴⁶³, gyrA⁹⁵, oxyR²⁸⁵, pncA⁵⁷ and the specific insertion element of *M. canettii* (IS canettii) allows the differentiation of *Mycobacterium* strains of *Mycobacterium* complex (see example 4).

10 The present invention provides an *in vitro* method for the detection and identification of *Mycobacteria* from the *Mycobacterium* complex in a biological sample, comprising the following steps:

- a) analysis for the presence or the absence of a nucleic acid fragment specifically deleted in *M. tuberculosis* excepted in strains of *M. tuberculosis* having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome, and
- 15 b) analysis of at least one additional genetic marker selected among RD1, RD2, RD3, RD4, RD5, RD6, RD7, RD8, RD9, RD10, RD11, RD13, RD14, RvD1, RvD2, RvD3, RvD4, RvD5, katG⁴⁶³, gyrA⁹⁵, oxyR²⁸⁵, pncA⁵⁷, the specific insertion element of *M. canettii*.

20 In a preferred embodiment, two additional markers are used, preferably RD4 and RD9. The analysis is performed by a technique selected among sequence hybridization, nucleic acid amplification, antigen-antibody complex.

It is also a goal of the present invention to provide a kit for the detection and identification of *Mycobacteria* from the *Mycobacterium* complex in a biological sample comprising the following elements:

- a) at least one pair of primers selected among nucleic acid fragments of the invention, and more preferably selected among the primers chosen from the group comprising SEQ ID N°13, SEQ ID N°14, SEQ ID N°15, SEQ ID N°16, SEQ ID N°17, SEQ ID N°18,
- 30 b) at least one pair of primers specific of the genetic markers selected among RD1, RD2, RD3, RD4, RD5, RD6, RD7, RD8, RD9, RD10, RD11, RD13, RD14, RvD1, RvD2, RvD3, RvD4, RvD5, katG⁴⁶³, gyrA⁹⁵, oxyR²⁸⁵, pncA⁵⁷, the specific insertion element of *M. canettii*.
- 35 c) the reagents necessary to carry out a DNA amplification reaction,

- d) optionally, the necessary components which make it possible to verify or compare the sequence and/or the size of the amplified fragment.

In a preferred embodiment, the kit comprises the following elements:

- 5 a) at least one pair of primers selected among nucleic acid fragments of the invention, and more preferably selected among the primers chosen from the group comprising SEQ ID N°13, SEQ ID N°14, SEQ ID N°15, SEQ ID N°16, SEQ ID N°17, SEQ ID N°18,
- b) one pair of primers specific of the genetic marker RD4,
- c) one pair of primers specific of the genetic marker RD9,
- 10 d) the reagents necessary to carry out a DNA amplification reaction,
- e) optionally, the necessary components which make it possible to verify or compare the sequence and/or the size of the amplified fragment.

15 The figures and examples presented below are provided as further guide to the practitioner of ordinary skill in the art and are not to be construed as limiting the invention in anyway.

FIGURES

20

Figure 1 : Amplicons obtained from strains that have the indicated genomic region present or deleted. Sizes of amplicons in each group are uniform. Numbers correspond to strain designation used in Kremer et al. (1999, J. Clin Microbiol. 37: 2607-2618) (Ref. 8) and Supply et al (2001, J. Clin. Microbiol. 39: 3563-3571) (ref.9).

25

Figure 2 : Sequences in the TbD1 region obtained from strains of various geographic regions.

* refers to groups based on *katG*^{c463}/*gyrA*^{c95} sequence polymorphism defined by Sreevatsan and colleagues (Ref. 2). Numbers correspond to strain designation used in Kremer et al. (1999, J. Clin Microbiol. 37: 2607-2618) (Ref. 8) and Supply et al (2001, J. Clin. Microbiol. 39: 3563-3571) (ref.9).

30

Figure 3 : Spoligotypes of selected *M. tuberculosis* and *M. bovis* strains. Numbers correspond to strain designation used in Kremer et al. (1999, J. Clin Microbiol. 37: 2607-2618) (Ref. 8) and Supply et al (2001, J. Clin. Microbiol. 39: 3563-3571) (ref.9).

- 5 **Figure 4 :** Scheme of the proposed evolutionary pathway of the tubercle bacilli illustrating successive loss of DNA in certain lineages (grey boxes). The scheme is based on presence or absence of conserved deleted regions and on sequence polymorphisms in five selected genes. Note that the distances between certain branches may not correspond to actual phylogenetic differences calculated by other methods.
- 10 Dark arrows indicate that strains are characterized by *katG*^{c463} CTG (Leu), *gyrA*^{c95} ACC (Thr), typical for group 1 organisms. Arrows with white lines indicate that strains belong to group 2 characterized by *katG*^{c463} CGG (Arg), *gyrA*^{c95} ACC (Thr). The arrow with white boxes indicates that strains belong to group 3, characterized by *katG*^{c463} CGG (Arg), *gyrA*^{c95} AGC (Ser), as defined by Sreevatsan and colleagues (Sreevastan et al., 1997 Proc. Natl. Acad.Sci USA 151: 9869-9874) (Ref. 2).
- 15

Figure 5 : Scheme of the TbD1 deletion and surrounding region in Mycobacterium complex.

A : Scheme of TbD1 and surrounding region in genome of *M. bovis*, *M. bovis* BCG, *M. africanum*, *M. canettii*, *M. microti* and ancestral strains of *M. tuberculosis* characterized by having the sequence CTG at codon 463 of gene *katG* and having no or very few IS6110 sequences inserted in their genome. The *mmpL6* gene, the *mmpS6* gene, the different primers, the different nucleic acid fragments and polypeptides coded by them are approximately localized in the region. The 2153 pb deletion named TbD1, specifically deleted in *M. tuberculosis* excepted in ancestral strains of *M. tuberculosis*, is delimited by its

20

25 two end points.

B : Scheme of TbD1 and surrounding region in genome of *M. tuberculosis* excepted ancestral strains of *M. tuberculosis*. Positions of the TbD1 deletion and of the nucleic acid of sequence SEQ ID N°1 in the genome of *M. tuberculosis* strain H37Rv are marked below the scheme. An chimeric ORF [*mmpS6-mmpL6*] resulting from the absence of TbD1 is drawn, the sequence of this chimeric ORF, SEQ ID N°21 and the sequence of the encoded polypeptide, SEQ ID N°22, are approximately localized above the scheme.

30

Figure 6 : Sequence of the specific insertion element in genome of *Mycobacterium canettii* strains. The beginning of this insertion element is at position 399 and the end of this insertion element is at position 2378. This insertion element contains the coding sequence of a

35

putative transposase (sequence in bold characters, from position 517 to position 2307) that shows significant homology with a transposase of *Mycobacterium smegmatis*. This coding sequence is framed by two 20 bp inverted repeats (sequences underlined from position 399 to 418 and from position 2359 to 2378).

5

EXAMPLES

1. MATERIAL AND METHODS:

10

1.1. Bacterial Strains: The 100 *M. tuberculosis* complex strains comprised 46 *M. tuberculosis* strains isolated in 30 countries, 14 *M. africanum* strains, 28 *M. bovis* strains originating in 5 countries, 2 *M. bovis* BCG vaccine strains (Pasteur and Japan), 5 *M. microti* strains, and 5 *M. canettii* strains. The strains were isolated from human and animal sources and were selected to represent a wide diversity including 60 strains that have been used in a multi-center study (8). The *M. africanum* strains were retrieved from the collection of the Wadsworth Center, New York State Department of Health, Albany, New York, whereas the majority of the *M. bovis* isolates came from the collection of the University of Zaragoza, Spain. Four *M. canettii* strains are from the culture collection of the Institut Pasteur, Paris, France. The strains have been extensively characterized by reference typing methods, i.e. IS6110 restriction fragment length polymorphism (RFLP) typing and spoligotyping. *M. tuberculosis* H37Rv, *M. tuberculosis* H37Ra, *M. tuberculosis* CDC1551, *M. bovis* AF2122/97, *M. microti* OV254, and *M. canettii* CIPT 140010059 were included as reference strains. DNA was prepared as previously described (10).

25

1.2. Genome comparisons and primer design

For preliminary genome comparisons between *M. tuberculosis* and *M. bovis* websites <http://genolist.pasteur.fr/TubercuList/> and http://www.sanger.ac.uk/Projects/M_bovis/ as well as inhouse databases were used. For primer design, sequences inside or flanking RD and RvD regions were obtained from the same websites. Primers were designed using the primer 3 website <http://www-genome.wi.mit.edu/cgi-bin/primer/primer3 WWW.cgi> that would amplify ca. 500 base pair fragments in the reference strains (Table 1).

30

1.3. RD-PCR analysis

Reactions were performed in 96 well plates and contained per reaction 1.25 µl of 10 x PCR buffer (600mM Tris HCl pH 8.8, 20 mM MgCl₂, 170 mM (NH₄)₂SO₄, 100 mM β-mercaptoethanol), 1.25 µl 20mM nucleotide mix, 50 nM of each primer, 1-10 ng of template DNA, 10% DMSO, 0.2 units *Taq* polymerase (Gibco-BRL) and sterile distilled water to 12.5 µl. Thermal cycling was performed on a PTC-100 amplifier (MJ Inc.) with an initial denaturation step of 90 seconds at 95°C, followed by 35 cycles of 30 seconds at 95°C, 1 min at 58°C, and 4 min at 72°C.

1.4. Sequencing of junction regions (RDs, TbD1,) *katG*, *gyrA*, *oxyR* and *pncA* genes

PCR products were obtained as described above, using primers listed in Table 1.

For primer elimination, 6 µl PCR product was incubated with 1 unit of Shrimp Alkaline phosphatase (USB), 10 units of exonuclease I (USB), and 2 µl of 5 x buffer (200mM Tris HCl pH 8.8, 5mM MgCl₂) for 15 min at 37°C and then for 15 min at 80°C. To this reaction mixture 2 µl of Big Dye sequencing mix (Applied Biosystems), 2 µl (2µM) of primer and 3 µl of 5 x buffer (5mM MgCl₂, 200mM Tris HCl pH 8.8) were added and 35 cycles (96°C for 30 sec; 56°C for 15 sec; 60°C for 4 min) performed in a thermocycler (MJ-research Inc., Watertown, MA). DNA was precipitated using 80 µl of 76% ethanol, centrifuged, rinsed with 70% ethanol, and dried. Reactions were dissolved in 2 µl of formamide/EDTA buffer, denatured and loaded onto 48 cm, 4 % polyacrylamide gels and electrophoresis performed on 377 automated DNA sequencers (Applied Biosystems) for 10 to 12 h. Alternatively, reactions were dissolved in 0.3 mM EDTA buffer and subjected to automated sequencing on a 3700 DNA sequencer (Applied Biosystems). Reactions generally gave between 500-700 bp of unambiguous sequence.

1.5. Accession Numbers

The sequence of the TbD1 region from the ancestral *M. tuberculosis* strain No. 74 (Ref. 8) containing genes *mmpS6* and *mmpL6* was deposited in the EMBL database under accession No. AJ426486. Sequences bordering RD4, RD7, RD8, RD9 and RD10 in BCG are available under accession numbers AJ003103, AJ007301, AJ131210, Y18604, and AJ132559, respectively.

2. EXPERIMENTAL DATA:

The distribution of 20 variable regions resulting from insertion-deletion events in the genomes of the tubercle bacilli has been evaluated in a total of 100 strains of *Mycobacterium tuberculosis*, *M. africanum*, *M. canettii*, *M. microti* and *M. bovis*. This approach showed that the majority of these polymorphisms did not occur independently in the different strains of the *M. tuberculosis* complex but, rather, result from ancient, irreversible genetic events in common progenitor strains. Based on the presence or absence of an *M. tuberculosis* specific deletion (TbD1), *M. tuberculosis* strains can be divided into ancestral and "modern" strains, the latter comprising representatives of major epidemics like the Beijing, Haarlem and African *M. tuberculosis* clusters. Furthermore, successive loss of DNA, reflected by RD9 and other subsequent deletions, was identified for an evolutionary lineage represented by *M. africanum*, *M. microti* and *M. bovis* that diverged from the progenitor of the present *M. tuberculosis* strains before TbD1 occurred. These findings contradict the often-presented hypothesis that *M. tuberculosis*, the etiological agent of human tuberculosis evolved from *M. bovis*, the agent of bovine disease. *M. canettii* and ancestral *M. tuberculosis* strains lack none of these deleted regions and therefore appear to be direct descendants of tubercle bacilli that existed before the *M. africanum*→*M. bovis* lineage separated from the *M. tuberculosis* lineage. This suggests that the common ancestor of the tubercle bacilli resembled *M. tuberculosis* or *M. canettii* and could well have been a human pathogen already.

The mycobacteria grouped in the *M. tuberculosis* complex are characterized by 99.9% similarity at the nucleotide level and identical 16S rRNA sequences (1, 2) but differ widely in terms of their host tropisms, phenotypes and pathogenicity. Assuming that they are all derived from a common ancestor, it is intriguing that some are exclusive human (*M. tuberculosis*, *M. africanum*, *M. canettii*) or rodent pathogens (*M. microti*) whereas others have a wide host spectrum (*M. bovis*). What was the genetic organization of the last common ancestor of the tubercle bacilli and in which host did it live? Which genetic events may have contributed to the fact that the host spectrum is so different and often specific? Where and when did *M. tuberculosis* evolve? Answers to these questions are important for a better understanding of the pathogenicity and the global epidemiology of tuberculosis and may help to anticipate future trends in the spread of the disease.

Because of the unusually high degree of conservation in their housekeeping genes it has been suggested that the members of the *M. tuberculosis* complex underwent an evolutionary bottleneck at the time of speciation, estimated to have occurred roughly 15,000 – 20,000 years ago (2). It also has been speculated that *M. tuberculosis*, the most widespread etiological agent of human tuberculosis has evolved from *M. bovis*, the agent of bovine

tuberculosis, by specific adaptation of an animal pathogen to the human host (3). However, both hypotheses were proposed before the whole genome sequence of *M. tuberculosis* (4) was available and before comparative genomics uncovered several variable genomic regions in the members of the *M. tuberculosis* complex. Differential hybridization arrays identified 14 regions (RD1 –14) ranging in size from 2 to 12.7 kb that were absent from BCG Pasteur relative to *M. tuberculosis* H37Rv (5, 6). In parallel, six regions, RvD1-5, and TbD1, that were absent from the *M. tuberculosis* H37Rv genome relative to other members of the *M. tuberculosis* complex were revealed by comparative genomics approaches employing pulsed-field gel electrophoresis (PFGE) techniques (5, 7) and *in silico* comparisons of the near complete *M. bovis* AF2122/97 genome sequence and the *M. tuberculosis* H37Rv sequence.

In the present study the inventors have analyzed the distribution of these 20 variable regions situated around the genome (Table 1) in a representative and diverse set of 100 strains belonging to the *M. tuberculosis* complex. The strains were isolated from different hosts, from a broad range of geographic origins, and exhibit a wide spectrum of typing characteristics like IS6110 and spoligotype hybridization patterns or variable-number tandem repeats of mycobacterial interspersed repetitive units (MIRU-VNTR) (8, 9). The inventors have found striking evidence that deletion of certain variable genomic regions did not occur independently in the different strains of the *Mycobacterium* complex and, assuming that there is little or no recombination of chromosomal segments between the various lineages of the complex, this allows the inventors to propose a completely new scenario for the evolution of the *Mycobacterium* complex and the origin of human tuberculosis.

Variable genomic regions and their occurrence in the members of the *M. tuberculosis* complex.

The PCR screening assay for the 20 variable regions (Table 1) within 46 *M. tuberculosis*, 14 *M. africanum*, 5 *M. canettii*, 5 *M. microti*, 28 *M. bovis* and 2 BCG strains employed oligonucleotides internal to known RDs and RvDs, as well as oligonucleotides flanking these regions (Table 1). This approach generated a large data set that was robust, highly reliable, and internally controlled since PCR amplicons obtained with the internal primer pair correlated with the absence of an appropriately sized amplicon with the flanking primer-pair, and *vice-versa*.

According to the conservation of junction sequences flanking the variable regions three types of regions were distinguished, each having different importance as an

evolutionary marker. The first type included mobile genetic elements, like the prophages phiRv1 (RD3) and phiRv2 (RD11) and insertion sequences IS1532 (RD6) and IS6110 (RD5), whose distribution in the tubercle bacilli was highly divergent (Table 2). The second type of deletion is mediated by homologous recombination between adjacent IS6110
 5 insertion elements resulting in the loss of the intervening DNA segment (RvD2, RvD3, RvD4, and RvD5 (7)) and is variable from strain to strain (Table 2).

The third type includes deletions whose bordering genomic regions typically do not contain repetitive sequences. Often this type of deletion occurred in coding regions resulting in the truncation of genes that are still intact in other strains of the *M. tuberculosis* complex.
 10 The exact mechanism leading to this type of deletion remains obscure, but possibly rare strand slippage errors of DNA polymerase may have contributed to this event. As shown in detail below, RD1, RD2, RD4, RD7, RD8, RD9, RD10, RD12, RD13, RD14, and TbD1 are representatives of this third group whose distribution among the 100 strains allows us to propose an evolutionary scenario for the members of the *M. tuberculosis* complex, that
 15 identified *M. tuberculosis* and/or *M. canettii* as most closely related to the common ancestor of the tubercle bacilli.

2.1. *M. tuberculosis* strains:

Investigation of the 46 *M. tuberculosis* strains by deletion analysis revealed that most
 20 RD regions were present in all *M. tuberculosis* strains tested (Table 2). Only regions RD3 and RD11, corresponding to the two prophages phiRv1 and phiRv2 of *M. tuberculosis* H37Rv (4), RD6 containing the insertion sequence IS1532, and RD5 that is flanked by a copy of IS6110 (5) were absent in some strains. This is an important observation as it implies that *M. tuberculosis* strains are highly conserved with respect to RD1, RD2, RD4, RD7,
 25 RD8, RD9, RD10, RD12, RD13, and RD14, and that these RDs represent regions that can differentiate *M. tuberculosis* strains independent of their geographical origin and their typing characteristics from certain other members of the *M. tuberculosis* complex. Furthermore, this suggests that these regions may be involved in the host specificity of *M. tuberculosis*.

In contrast, the presence or absence of RvD regions in *M. tuberculosis* strains was
 30 variable. The region which showed the greatest variability was RvD2, since 18 from 46 tested *M. tuberculosis* strains did not carry the RvD2 region. Strains with a high copy number of IS6110 (>14) missed regions RvD2 to RvD5 more often than strains with only a few copies. As an example, all six tested strains belonging to the Beijing cluster (8) lacked regions RvD2 and RvD3. This is in agreement with the proposed involvement of
 35 recombination of two adjacent copies of IS6110 in this deletion event (7).

However, the most surprising finding concerning the RvD regions was that TbD1 was absent from 40 of the tested *M. tuberculosis* strains (87 %), including representative strains from major epidemics such as the Haarlem, Beijing and Africa clusters (8). To accentuate this result we named this region "*M. tuberculosis* specific deletion 1" (TbD1). *In silico* sequence comparison of *M. tuberculosis* H37Rv with the corresponding section in *M. bovis* AF2122/97 revealed that in *M. bovis* this locus comprises two genes encoding membrane proteins belonging to a large family, whereas in *M. tuberculosis* H37Rv one of these genes (*mmpS6*) was absent and the second was truncated (*mmpL6*). Unlike the RvD2-RvD5 deletions, the TbD1 region is not flanked by a copy of IS6110 in *M. tuberculosis* H37Rv, suggesting that insertion elements were not involved in the deletion of the 2153 bp fragment. To further investigate whether the 40 *M. tuberculosis* strains lacking the TbD1 region had the same genomic organization of this locus as *M. tuberculosis* H37Rv, we amplified the TbD1-junction regions of the various strains by PCR using primers flanking the deleted region (Table 1). This approach showed that the size of the amplicons obtained from multiple strains was uniform (Fig. 1) and subsequent sequence analysis of the PCR products revealed that in all tested TbD1-deleted strains the sequence of the junction regions was identical to that of *M. tuberculosis* H37Rv (Fig.2). The perfect conservation of the junction sequences in TbD1-deleted strains of wide geographical diversity suggests that the genetic event which resulted in the deletion occurred in a common progenitor. However, six *M. tuberculosis* strains, all characterized by very few or no copies of IS6110 and spoligotypes that resembled each other (Fig. 3) still had the TbD1 region present. Interestingly, these six strains were also clustered together by MIRU-VNTR analysis (9).

Analysis of partial gene sequences of *oxyR*, *pncA*, *katG*, and *gyrA* which have been described as variable between different tubercle bacilli (2, 11, 12, 13) revealed that all tested *M. tuberculosis* strains showed *oxyR* and *pncA* partial sequences typical for *M. tuberculosis* (*oxyR* - nucleotide 285 (*oxyR*²⁸⁵):G, *pncA* - codon 57 (*pncA*⁵⁷: CAC). Based on the *katG* codon 463 (*katG*⁴⁶³) and *gyrA* codon 95 (*gyrA*⁹⁵) sequence polymorphism, Sreevatsan and colleagues (2) defined three groups among the tubercle bacilli, group 1 showing *katG*⁴⁶³ CTG (Leu), *gyrA*⁹⁵ ACC (Thr), group 2 exhibiting *katG*⁴⁶³ CGG (Arg), *gyrA*⁹⁵ ACC (Thr), and group 3 showing *katG*⁴⁶³ CGG (Arg), *gyrA*⁹⁵ AGC (Ser). According to this scheme, in our study 16 of the 46 tested *M. tuberculosis* strains belonged to group 1, whereas 27 strains belonged to group 2 and only 3 isolates to group 3. From the 40 strains that were deleted for region TbD1, 9 showed characteristics of group 1, including the strains belonging to the Beijing cluster, 28 of group 2, including the strains from the Haarlem and Africa clusters and 3 of group 3, including H37Rv and H37Ra. Most interestingly, all six *M. tuberculosis* strains

where the TbD1 region was not deleted, contained a leucine (CTG) at *katG*⁴⁶³, which was described as characteristic for ancestral *M. tuberculosis* strains (group 1) (2). As shown in Figure 4, this suggests that during the evolution of *M. tuberculosis* the *katG* mutation at codon 463 CTG (Leu) → CGG (Arg) occurred in a progenitor strain that had region TbD1 deleted. This proposal is supported by the finding that strains belonging to group 1 may or may not have deleted region TbD1, whereas all 30 strains belonging to groups 2 and 3 lacked TbD1 (Fig. 4). Furthermore, all strains of groups 2 and 3 characteristically lacked spacer sequences 33-36 in the direct repeat (DR) region (Fig. 3). It appears that such spacers may be lost but not gained (14). Therefore, TbD1 deleted strains will be referred to hereafter as “modern” *M. tuberculosis* strains.

2.2. *M. canettii*:

M. canettii is a very rare smooth variant of *M. tuberculosis*, isolated usually from patients from, or with connection to, Africa. Although it shares identical 16S rRNA sequences with the other members of the *Mycobacterium* complex, *M. canettii* strains differ in many respects including polymorphisms in certain house-keeping genes, IS1081 copy number, colony morphology, and the lipid content of the cell wall (15, 16). Therefore, we were surprised to find that in *M. canettii* all the RD, RvD, and TbD1 regions except the prophages (phiRv1, phiRv2) were present. In contrast, we identified a region (RD^{can}) being specifically absent from all five *M. canettii* strains that partially overlapped RD12 (Fig. 4).

The conservation of the RD, RvD, and TbD1 regions in the genome of *M. canettii* in conjunction with the many described and observed differences suggest that *M. canettii* diverged from the common ancestor of the *Mycobacterium* complex before RD, RvD and TbD1 occurred in the lineages of tubercle bacilli (Fig. 4). This hypothesis is supported by the finding that *M. canettii* was shown to carry 26 unique spacer sequences in the direct repeat region (14), that are no longer present in any other member of the *Mycobacterium* complex. An other specific feature of *M. canettii* is the presence of an insertion element whose sequence has been searched, by using PCR and hybridization approaches, without success in the other member strains of *Mycobacterium* complex (including *M. tuberculosis*, *M. bovis*, *M. africanum* and *M. microti*). This insertion element contained an ORF encoding a putative transposase framed by two inverted repeats. The sequence of this insertion element is represented in figure 6 and in SEQ ID N°19 where it begins at position 399 and ends at position 2378. The amino acids sequence of the putative transposase is drawn in SEQ ID N°20. As such, this insertion element can be used to differentiate between *M. tuberculosis* ancestral strains and *M. canettii* strains that may show the same TbD1, RD4 and RD9

profiles. Therefore, *M. canettii* represents a fascinating tubercle bacillus, whose detailed genomic analysis may reveal further insights into the evolution of *Mycobacterium* complex.

2.3. *M. africanum*:

5 The isolates designated as *M. africanum* studied here originate from West and East-African sources. 11 strains were isolated in Sierra Leone, Nigeria and Guinea and 2 strains in Uganda. One strain comes from the Netherlands.

For the 11 West African isolates, RD analysis indicated that these strains all lack the RD9 region containing *cobL*. Sequence analysis of the RD9 junction region showed that the
10 genetic organization of this locus in West African strains was identical to that of *M. bovis* and *M. microti* in that the 5' part of *cobL* as well as the genes Rv2073c and Rv2074c were absent. In addition, six strains (2 from Sierra Leone, 4 from Guinea) also lacked RD7, RD8 and RD10 (Table 2). The junction sequences bordering RD7, RD8 and RD10, like those for RD9, were identical to those of *M. bovis* and *M. microti* strains. As regards the two
15 prophages phiRv1 and phiRv2, the West African strains all contained phiRv2, whereas phiRv1 was absent. No variability was seen for the RvD regions. RvD1-RvD5 and TbD1 were present in all tested West African strains. This shows that *M. africanum* prevalent in West Africa can be differentiated from "modern" *M. tuberculosis* by at least two variable genetic markers, namely the absence of region RD9 and the presence of region TbD1.

20 In contrast, for East African *M. africanum* and for the isolate from the Netherlands, no genetic marker was found which could differentiate them from *M. tuberculosis* strains. With the exception of prophage phiRv1 (RD3) the 3 strains from Uganda and the Netherlands did not exhibit any of the RD deletions, but lacked the TbD1 region, as do "modern" *M. tuberculosis* strains. The absence of the TbD1 region was also confirmed by
25 sequence analysis of the TbD1 junction region, which was found to be identical to that of TbD1 deleted *M. tuberculosis* strains. These results indicate a very close genetic relationship of these strains to *M. tuberculosis* and suggest that they should be regarded as *M. tuberculosis* rather than *M. africanum* strains.

2.4. *M. microti*:

30 *M. microti* strains were isolated in the 1930's from voles (17) and more recently from immuno-suppressed patients (18). These strains are characterized by an identical, characteristic spoligotype, but differ in their IS6110 profiles. Both, the vole and the human isolates, lacked regions RD7, RD8, RD9, and RD10 as well as a region that is specifically
35 deleted from *M. microti* (RD^{mic}). RD^{mic} was revealed by a detailed comparative genomics

study of *M. microti* isolates (19) using clones from a *M. microti* Bacterial Artificial Chromosome (BAC) library. RD^{mic} partially overlaps RD1 from BCG (data not shown). Furthermore, vole isolates missed part of the RD5 region, whereas this region was present in the human isolate. As the junction region of RD5 in *M. microti* was different to that in BCG
5 (data not shown), RD5 was not used as an evolutionary marker.

2.5. *M. bovis* and *M. bovis* BCG:

M. bovis has a very large host spectrum infecting many mammalian species, including man. The collection of *M. bovis* strains that was screened for the RD and RvD
10 regions consisted of 2 BCG strains and 18 "classical" *M. bovis* strains generally characterized by only one or two copies of IS6110 from bovine, llama and human sources in addition to three goat isolates, three seal isolates, two oryx isolates, and two *M. bovis* strains from humans that presented a higher number of IS6110 copies.

Excluding prophages, the distribution of RDs allowed us to differentiate five main
15 groups among the tested *M. bovis* strains. The first group was formed by strains that lack RD7, RD8, RD9, and RD10. Representatives of this group are three seal isolates and two human isolates containing between three and five copies of IS6110 (data not shown). Two oryx isolates harboring between 17 and 20 copies of IS6110 formed the second group that lacked parts of RD5 in addition to RD7-RD10, and very closely resembled the *M. microti*
20 isolates. However, they did not show RD^{mic}, the deletion characteristic of *M. microti* strains (data not shown). Analysis of partial *oxyR* and *pncA* sequences from strains belonging to groups one and two, showed sequence polymorphisms characteristic of *M. tuberculosis* strains (*oxyR*²⁸⁵: G, *pncA*⁵⁷: CAC, Ref. 12, 13).

Group three consists of goat isolates that lack regions RD5, RD7, RD8, RD9, RD10,
25 RD12, and RD13. As previously described by Aranaz and colleagues, these strains exhibited an adenosine at position 285 of the *oxyR* pseudogene that is specific for "classical" *M. bovis* strains whereas the sequence of the *pncA*⁵⁷ polymorphism was identical to that in *M. tuberculosis* (20). This is in good agreement with our results from sequence analysis (Table
30 2) and the finding that except for RD4, the goat isolates displayed the same deletions as "classical" *M. bovis* strains. Taken together, this suggests that the *oxyR*²⁸⁵ mutation (G → A) occurred in *M. bovis* strains before RD4 was lost. Interestingly, the most common *M. bovis* strains ("classical" *M. bovis* (21)), isolated from cattle from Argentina, the Netherlands, the UK and Spain, as well as from humans (e. g. multi-drug resistant *M. bovis*
35 from Spain (22)) showed the greatest number of RD deletions and appear to have undergone

the greatest loss of DNA relative to other members of the *M. tuberculosis* complex. These lacked regions RD4, RD5, RD6, RD7, RD8, RD9, RD10, RD12 and RD13, confirming results obtained with reference strains (5, 6). These strains together with the two BCG strains were the only ones that showed the *pncA*⁵⁷ polymorphism GAC (Asp) in addition to the
 5 *oxyR*²⁸⁵ mutation (G → A) characteristic of *M. bovis*. Analysis of BCG strains indicate that BCG lacked the same RD regions as “classical” *M. bovis* strains in addition to RD1, RD2 and RD14 which apparently occurred during and after the attenuation process (Fig. 4) (6, 23).

In contrast to RDs, the RvD regions were highly conserved in the *M. bovis* strains.
 10 With the exception of the two IS6110-rich oryx isolates, that lacked RvD2, RvD3 and RvD4, all other strains had the five RvD regions present. It is particularly noteworthy that TbD1 was present in all *M. bovis* strains.

However, except for the two human isolates, containing between three and five copies of IS6110 from group 1, strains designated as *M. bovis* showed a single nucleotide
 15 polymorphism in the TbD1 region at codon 551 (AAG) of the *mmpL6* gene, relative to *M. canettii*, *M. africanum* and ancestral *M. tuberculosis* strains, which are characterized by codon AAC. Even the strains isolated from seals and from oryx with *oxyR* or *pncA* loci like those of *M. tuberculosis* and with fewer deleted regions than the classical *M. bovis* strains, showed the *mmpL6*⁵⁵¹AAG polymorphism typical for *M. bovis* and *M. microti* (Table 2, Fig.
 20 4). As such, this polymorphism could serve as a very useful genetic marker for the differentiation of strains that lack RD7, RD8, RD9, and RD10 and have been classified as *M. bovis* or *M. africanum*, but may differ from other strains of the same taxon.

3. DISCUSSION

25

3.1. Origin of human tuberculosis

For many years, it was thought that human tuberculosis evolved from the bovine disease by adaptation of an animal pathogen to the human host (3). This hypothesis is based on the property of *M. tuberculosis* to be almost exclusively a human pathogen, whereas
 30 *M. bovis* has a much broader host range. However, the results from this study unambiguously show that *M. bovis* has undergone numerous deletions relative to *M. tuberculosis*. This is confirmed by the preliminary analysis of the near complete genome sequence of *M. bovis* AF2122/97, a “classical” *M. bovis* strain isolated from cattle, which revealed no new gene clusters that were confined specifically to *M. bovis*. This indicates that the genome of *M.*
 35 *bovis* is smaller than that of *M. tuberculosis* (24). It seems plausible that *M. bovis* is the final

member of a separate lineage represented by *M. africanum* (RD9), *M. microti* (RD7, RD8, RD9, RD10) and *M. bovis* (RD4, RD5, RD7, RD8, RD9, RD10, RD12, RD13) (25) that branched from the progenitor of *M. tuberculosis* isolates. Successive loss of DNA may have contributed to clonal expansion and the appearance of more successful pathogens in certain new hosts.

Whether the progenitor of extant *M. tuberculosis* strains was already a human pathogen when the *M. africanum* → *M. bovis* lineage separated from the *M. tuberculosis* lineage is a subject for speculation. However, we have two reasons to believe that this was the case. Firstly, the six ancestral *M. tuberculosis* strains (TbD1⁺, RD9⁺) (Fig.3) that resemble the last common ancestor before the separation of *M. tuberculosis* and *M. africanum* are all human pathogens. Secondly, *M. canettii*, which probably diverged from the common ancestor of today's *M. tuberculosis* strains prior to any other known member of the *M. tuberculosis* complex is also a human pathogen. Taken together, this means that those tubercle bacilli, which are thought to most closely resemble the progenitor of *M. tuberculosis* are human and not animal pathogens. It is also intriguing that most of these strains were of African or Indian origin (Fig. 3). It is likely that these ancestral strains predominantly originated from endemic foci (15, 26), whereas "modern" *M. tuberculosis* strains that have lost TbD1 may represent epidemic *M. tuberculosis* strains that were introduced into the same geographical regions more recently as a consequence of the worldwide spread of the tuberculosis epidemic.

3.2. The evolutionary timescale of the *M. tuberculosis* complex

Because of the high sequence conservation in housekeeping genes, Sreevatsan *et al.* previously hypothesized that the tubercle bacilli encountered a major bottleneck 15,000 – 20,000 years ago (2). As the conservation of the TbD1 junction sequence in all tested TbD1 deleted strains suggests descentance from a single clone, the TbD1 deletion is a perfect indicator that "modern" *M. tuberculosis* strains that account for the vast majority of today's tuberculosis cases definitely underwent such a bottleneck and then spread around the world.

As described in detail in the results section, our analysis showed that the *katG*⁴⁶³ CTG→CGG and the subsequent *gyrA*⁹⁵ ACC →AGC mutations, that were used by Sreevatsan and colleagues to designate groups 2 and 3 of their proposed evolutionary pathway of the tubercle bacilli (2), occurred in a lineage of *M. tuberculosis* strains that had already lost TbD1 (Fig.4). Although deletions are more stable markers than point mutations, which may be subject to reversion, a perfect correlation of deletion and point mutation data was found for the tested strains.

This information, together with results from a recent study by Fletcher and colleagues (27), who have shown that *M. tuberculosis* DNAs amplified from naturally mummified Hungarian villagers from the 18th and 19th century belonged to *katG*⁴⁶³/*gyrA*⁹⁵ groups 2 and 3, suggests that the TbD1 deletion occurred in the lineage of *M. tuberculosis* before the 18th century. This could mean that the dramatic increase of tuberculosis cases later in the 18th century in Europe mainly involved “modern” *M. tuberculosis* strains. In addition, it shows that tuberculosis was caused by *M. tuberculosis* and not by *M. bovis*, a fact which is also described for cases in rural medieval England (28).

There is good evidence that mycobacterial infections occurred in man several thousand years ago. We know that tuberculosis occurred in Egypt during the reign of the pharaohs because spinal and rib lesions pathognomonic of tuberculosis have been identified in mummies from that period (29). Identification of acid fast bacilli as well as PCR amplification of IS6110 from Peruvian mummies (30) also suggest that tuberculosis existed in pre-Columbian societies of Central and South America. To estimate when the TbD1 bottleneck occurred, it would now be very interesting to know whether the Egyptian and South American mummies carried *M. tuberculosis* DNA that had TbD1 deleted or not.

The other major bottleneck, which seems to have occurred for members of the *M. africanum* → *M. microti* → *M. bovis* lineage is reflected by RD9 and the subsequent RD7, RD8 and RD10 deletions (Fig. 4). These deletions seem to have occurred in the progenitor of tubercle bacilli that - today - show natural host spectra as diverse as humans in Africa, voles on the Orkney Isles (UK), seals in Argentina, goats in Spain, and badgers in the UK. For this reason it is difficult to imagine that spread and adaptation of RD9-deleted bacteria to their specific hosts could have appeared within the postulated 15,000 – 20,000 years of speciation of the *M. tuberculosis* complex.

However, more insight into this matter could be gained by RD analysis of ancient DNA samples, e. g. mycobacterial DNA isolated from a 17,000 year old bison skeleton (31). The mycobacterium whose DNA was amplified showed a spoligotype that was most closely related to patterns of *M. africanum* and could have been an early representative of the lineage *M. africanum* → *M. bovis*. With the TbD1 and RD9 junction sequences that we supply here, PCR analyses of ancient DNAs should enable very focused studies to be undertaken to learn more about the timescale within which the members of the *M. tuberculosis* complex have evolved.

3.3. Concluding comments

Our study provides an overview of the diversity and conservation of variable regions

in a broad range of tubercle bacilli. Deletion analysis of 100 strains from various hosts and countries has identified some evolutionarily "old" *M. canettii*, *M. tuberculosis* and *M. africanum* strains, most of them of African origin, as well as "modern" *M. tuberculosis* strains, the latter including representatives from major epidemic clusters like Beijing, Haarlem and Africa. The use of deletion analysis in conjunction with molecular typing and analysis of specific mutations was shown to represent a very powerful approach for the study of the evolution of the tubercle bacilli and for the identification of evolutionary markers. In a more practical perspective, these regions, primarily RD9 and TbD1 but also RD1, RD2, RD4, RD7, RD8, RD10, RD12 and RD13 represent very interesting candidates for the development of powerful diagnostic tools for the rapid and unambiguous identification of members of the *M. tuberculosis* complex (32). This genetic approach for differentiation can now be used to replace the often confusing traditional division of the *M. tuberculosis* complex into rigidly defined subspecies.

Moreover, functional analyses will show whether the TbD1 deletion confers some selective advantage to "modern" *M. tuberculosis*, or whether other circumstances contributed to the pandemic of the TbD1 deleted *M. tuberculosis* strains.

EXAMPLE 4

The members of the *M. tuberculosis* complex share an unusually high degree of conservation such that the commercially-available nucleic acid probes and amplification assays cannot differentiate these organisms. In addition conventional identification methods are often ambiguous, cumbersome and time consuming because of the slow growth of the organisms.

In the present invention the inventors, by a deletion analysis, solve the problem faced by clinical mycobacteriology laboratories for differentiation within the *M. tuberculosis* complex.

This approach allows to perform a diagnostic on a biological fluid by using at least three markers including TbD1. The following table 3 illustrates such a combinaison sufficient to realize the distinction between the members of the *Mycobacterium* complex.

MYCOBACTERIUM STRAIN	MARKERS		
	RD4	RD9	TbD1
<i>M. bovis</i> BCG	-	-	+
<i>M. bovis</i>	-	-	+
<i>M. africanum</i>	+	-	+
<i>M. tuberculosis</i>	+	+	-
<i>M. tuberculosis</i> ancestral	+	+	+
<i>M. canettii</i>	+	+	+

Table 3

Beside TbD1 marker, preferably at least 2 other markers should be used. Examples of such additional markers available in the literature are listed in the following table 1.

- 5 Although ancestral strains of *Mycobacterium tuberculosis* represent only 5% of all *Mycobacterium tuberculosis* strains, persons who would be interested in distinguishing the ancestral strains of *Mycobacterium tuberculosis* from the strains of *Mycobacterium canettii*, could consider using the genetic marker RD12 in combination with the three markers described in table 3. Because the region RD^{can} partially overlapped RD12 in genome of
- 10 *Mycobacterium canettii*, flanking primers as described in table 1 do not hybridize on genomic DNA of *Mycobacterium canettii*. Therefore, PCR amplification with these flanking

primers results in 2.8 kb PCR product in *Mycobacterium tuberculosis* and no PCR product in *Mycobacterium canettii*.

An other way to distinguish ancestral strains of *Mycobacterium tuberculosis* from *Mycobacterium canettii* would be the detection of the insertion element specific for *M.*

5 *canettii* strains and corresponding to SEQ ID N° 19.

Supplemental data:

10

Table 1: RD, RvD and TbD1 regions and selected primers

Region absent from BCG	Gene	Size (kb)	Internal Primerpair	Flanking primers or 2 nd internal * primerpair
RD1	Rv3871-Rv3879c	9.5	RD1in-Rv3878F GTC AGC CAA GTC AGG CTA CC RD1in-Rv3878R CAA CGT TGT GGT TGT TGA GG	RD1-flank.left GAA ACA GTC CCC AGC AGG T RD1-flank.right TTC AAC GGG TTA CTG CGA AT
RD2	Rv1978-Rv1988	10.8	RD2-Rv1979.int.F TAT AGC TCT CGG CAG GTT CC RD2-Rv1979-int.R ATC GGC ATC TAT GTC GGT GT	RD2-flank.F CTC GAC CGC GAC GAT GTG C RD2-flank.R CCT CGT TGT CAC CGC GTA TG
RD3*	Rv1573-Rv1586c	9.2	RD3-Rv1586.int.F TTA TCT TGG CGT TGA CGA TG RD3-Rv1586.int.R CAT ATA AGG GTG CCC GCT AC	RD3-int-REP.F CTG ACG TCG TTG TCG AGG TA* RD3-int-REP.R GTA CCC CCA GGC GAT CTT*
RD4	Rv1505c-Rv1516c	12.7	RD4-Rv1516.int.F CAA GGG GTA TGA GGT TCA CG RD4-Rv1516.int.R CGG TGA TTC GTG ATT GAA CA	RD4-flank.F CTC GTC GAA GGC CAC TAA AG RD4-flank.R AAG GCG AAC AGA TTC AGC AT

Table 1 (continued)

RD5*	Rv2346c-Rv2353c	9.0	RD5A-Rv2348.int.F AAT CAC GCT GCT GCT ACT CC RD5A-Rv2348.int.R GTG CTT TTG CCT CTT GGT C	RD5B-plcA.int.F CAA GTT GGG TCT GGT CGA AT RD5B-plcA.int.R GCT ACC CAA GGT CTC CTG GT
RD6*	Rv3425-Rv3428c	4.9	RD6-IS1532F CAG CTG GTG AGT TCA AAT GC RD6-IS1532R CTC CCG ACA CCT GTT CGT	ND ND
RD7	Rv1964-Rv1977	12.7	RD7-Rv1976.int.F TGG ATT GTC GAC GGT ATG AA RD7-Rv1976.int.R GGT CGA TAA GGT CAC GGA AC	RD7-flank.F GGT AAT CGT GGC CGA CAA G RD7-flank.R CAG CTC TTC CCC TCT CGA C
RD8	<i>ephA-lpqG</i>	5.9	RD8-ephA.F GGT GTG ATT TGG TGA GAC GAT G RD8-ephA.R AGT TCC TCC TGA CTA ATC CAG GC	RD8-flank.F CAA TCA GGG CTG TGC TAA CC RD8-flank.R CGA CAG TTG TGC GTA CTG GT
RD9	<i>cobL</i> -Rv2075	2.0	RD9-intF CGA TGG TCA ACA CCA CTA CG RD9-intR CTG GAC CTC GAT GAC CAC TC	RD9-flankF GTG TAG GTC AGC CCC ATC C RD9-flankR GCC CAA CAG CTC GAC ATC
RD10	Rv0221-Rv0223	1.9	RD10-intF GTA ACC GCT TCA CCG GAA T RD10-intR GTC AAC TCC ACG GAA AGA CC	RD10-flankF CTG CAA CCA TCC GGT ACA C RD10-flankR GTC ATG AAC GCC GGA CAG
RD11	Rv2645-Rv2695c	11.0	RD11-Rv2646F CGG CAG CTA GAC GAC CTC RD11-Rv2646R AAC GTG CTG CGA TAG GTT TT	RD11-fla-F TCA CAT AGG GGC TGC GAT AG RD11-fla-R AGA GGA ACC TTT CGG TGG TT
RD12	<i>sseC</i> -Rv3121	2.8	RD12-Rv3120.int.F GAA ATA CGA GTG CGC TGA CC RD12-Rv3120.int.R CTC TGA ACC ATC GGT GTC G	RD12-flank.F GCC ATC AAC GTC AAG AAC CT RD12-flank.R CGG CCA GGT AAC AAG GAG T
RD13	Rv1255c-Rv1257c	3.0	RD13intF GGA TGT CAC TCG GAA CGG CA RD13intR CAC CGG GCT GAT CGA GCG A	RD13-flank.F CGA TGG TGT TTC TTG GTG AG RD13-flank.R GGA TCG GCT CAG TGA ATA CC
RD14	Rv1765c-Rv1773c	9.0	RD14-Rv1769.int.F GTG GAG CAC CTT GAC CTG AT RD14-Rv1769.int.R CGT CGA ATA CGA GTC GAA CA	RD14-flankF TTG ATT CGC CAA CAA CTG AA RD14-flankR GGG CTG GTT AGT GTC GAT TC

Table 1 (continued)Region missing from *M. tuberculosis* H37Rv

RvD1*		5.0	RvD1-int1F AGC GCG TCG AAC ACC GGC RvD1-int1R CCT GAA TCC GCG CAA TTC CAT	RvD1-int2.F GAG CCA CTC CGA TGT TGA CT RvD1-int2.R CAC GCG AAC CCT ACC TAC AT
RvD2*	<i>plcD</i>	5.1	RvD2-int1F GTT CTC CTG TCG AAC CTC CA RvD2-int1R ACT TCA CCG GTT TCA TCT CG	RvD2-int2F GGA CGG TGA CGG TAT TTG TC RvD2-int2R TCG CCA ACT TCT ATG GAC CT
RvD3		1.0	RvD3-intF ATC GAT CAG GTC GTC AAT GC RvD3-intR ACG CCA CCA TCA AGA TCC	RvD3-flank.F AAA CCA TGC AGC GTC TGC CA RvD3-flankR GCG TTT CTG CGT CTG GTT GA
RvD4*	PPE gene	0.8	RvD4-intF-PPE GGT TGC CAA CGT TAC CGA TGC RvD4-intR-PPE CCG GTG GTG GTG GCG GCT	ND ND
RvD5	<i>moa</i>	4.0	RvD5intF GGG TTC ACG TTC ATT ACT GTT C RvD5intR CCT GCG CTT ATC TCT AGC GG	RvD5-flankF CCC ATC GTG GTC GTT CAC C RvD5-flankR GTA CCC GCA CCA CCT GCT G
TbD1	<i>mmpL6</i>	2.1	TbD1intS.F CGT TCA ACC CCA AAC AGG TA TbD1intS.R AAT CGA ACT CGT GGA ACA CC	TbD1fla1-F CTA CCT CAT CTT CCG GTC CA TbD1fla1-R CAT AGA TCC CGG ACA TGG TG
<i>katG</i>, <i>gyrA</i>, <i>oxyR</i>', <i>pncA</i> and <i>mmpL6</i> PCR and sequencing primers				
<i>katG</i> ⁴⁶³			<i>katG</i> -2154,225-PCR-F CTA CCA GCA CCG TCA TCT CA <i>katG</i> -2155,157-PCR-R AGG TCG TAT GGA CGAACA CC	<i>katG</i> -2154,872-SEQ-R ACA AGC TGA TCC ACC GAG AC
<i>gyrA</i> ⁹⁵			<i>gyrA</i> -7,127-PCR-F GTT CGT GTG TTG CGT CAA GT <i>gyrA</i> - 8,312-PCR-R CAG CTG GGT GTG CTT GTA AA	<i>gyrA</i> -7,461F CGG GTG CTC TAT GCA ATG TT
<i>oxyR</i> ²⁸⁵			<i>oxyR</i> 2725,559F TAT GCG ATC AGG CGT ACT TG <i>oxyR</i> -2726,024-PCR-R CAA AGC AGT GGT TCA GCA GT	<i>oxyR</i> -2726,024-SEQ-R CAA AGC AGT GGT TCA GCA GT

Table 1 (continued)

<i>pncA</i> ⁵⁷	<i>pncA</i> -2288,678-PCR-F	<i>pncA</i> - 2289,319-SEQ-R
	ATC AGG AGC TGC AAA CCA AC	GGC GTC ATG GAC CCT ATA TC
<i>mmpL6</i> ⁵⁵¹	<i>pncA</i> - 2289,319-PCR-R	
	GGC GTC ATG GAC CCT ATA TC	
	<i>mmpL</i> -seq5F	<i>mmpL</i> -seq5F
	GTA TCA GAG GGA CCG AGC AG	GTA TCA GAG GGA CCG AGC AG
	TBD1fla1-R	
	CAT AGA TCC CGG ACA TGG TG	

The RD nomenclature used in this table is based on that used by Brosch *et al.* (2000), (Ref. 25) and differs from that proposed by Behr and coworkers (1999), (Ref. 6). Primer sequences are shown in 5' →3' direction.

* Regions where a second pair of internal primers was used rather than flanking primers, due to

5 flanking repetitive regions, and/or mobile genetic elements.

REFERENCES

1. Boddingtonhaus, B., Rogall, T., Flohr, T., Blocker, H. & Bottger, E. C. (1990) *J Clin Microbiol* **28**, 1751-9.
2. Sreevatsan, S., Pan, X., Stockbauer, K. E., Connell, N. D., Kreiswirth, B. N., Whittam, T. S. & Musser, J. M. (1997) *Proc Natl Acad Sci USA* **94**, 9869-74.
3. Stead, W. W., Eisenach, K. D., Cave, M. D., Beggs, M. L., Templeton, G. L., Thoen, C. O. & Bates, J. H. (1995) *Am J Respir Crit Care Med* **151**, 1267-8.
4. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., 3rd, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Barrell, B. G. & et al. (1998) *Nature* **393**, 537-44.
5. Gordon, S. V., Brosch, R., Billault, A., Garnier, T., Eiglmeier, K. & Cole, S. T. (1999) *Mol Microbiol* **32**, 643-55.
6. Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S. & Small, P. M. (1999) *Science* **284**, 1520-3.
7. Brosch, R., Philipp, W. J., Stavropoulos, E., Colston, M. J., Cole, S. T. & Gordon, S. V. (1999) *Infect Immun* **67**, 5768-74.

8. Kremer, K., van Soolingen, D., Frothingham, R., Haas, W. H., Hermans, P. W., Martin, C., Palittapongarnpim, P., Plikaytis, B. B., Riley, L. W., Yakrus, M. A., Musser, J. M. & van Embden, J. D. (1999) *J Clin Microbiol* **37**, 2607-18.
9. Supply, P., Lesjean, S., Savine, E., Kremer, K., van Soolingen, D., & Locht, C.
5 (2001) *J Clin Microbiol* **39**, 3563-71.
10. Van Soolingen, D., de Haas, P. E. W., Hermans, P. W. M. & van Embden, J. D. A. (1994) *Methods Enzymol* **235**, 196-205.
11. Heym, B., Honore, N., Truffot-Pernot, C., Banerjee, A., Schurra, C., Jacobs, W. R., Jr., van Embden, J. D., Grosset, J. H. & Cole, S. T. (1994) *Lancet* **344**, 293-8.
- 10 12. Scorpio, A., Collins, D., Whipple, D., Cave, D., Bates, J. & Zhang, Y. (1997) *J Clin Microbiol* **35**, 106-10.
13. Sreevatsan, S., Escalante, P., Pan, X., Gillies, D. A., 2nd, Siddiqui, S., Khalaf, C. N., Kreiswirth, B. N., Bifani, P., Adams, L. G., Ficht, T., Perumaalla, V. S., Cave, M. D., van Embden, J. D. & Musser, J. M. (1996) *J Clin Microbiol* **34**, 2007-10.
- 15 14. Van Embden, J. D., van Gorkom, T., Kremer, K., Jansen, R., van Der Zeijst, B. A. & Schouls, L. M. (2000) *J Bacteriol* **182**, 2393-401.
15. Van Soolingen, D., Hoogenboezem, T., de Haas, P. E., Hermans, P. W., Koedam, M. A., Teppema, K. S., Brennan, P. J., Besra, G. S., Portaels, F., Top, J., Schouls, L. M. & Van Embden, J. D. (1997) *Int J Syst Bacteriol* **47**, 1236-45.
- 20 16. Papa, F., Laszlo, A., David, H. L. & Daffe, M. (1989) *Acta Leprol* **7** (Suppl.) 98-101.
17. Wells, A. Q., (1937) *Lancet* 1221.
18. Van Soolingen, D., Van der Zanden, A. G., de Haas, P. E., Noordhoek, G. T., Kiers, A., Foudraïne, N. A., Portaels, F., Kolk, A. H., Kremer, K. & Van Embden, J. D. (1998) *J Clin Microbiol* **36**, 1840-5.
- 25 19. Brodin, P., *et al.* (2002) in preparation
20. Aranaz, A., Liebana, E., Gomez-Mampaso, E., Galan, J. C., Cousins, D., Ortega, A., Blazquez, J., Baquero, F., Mateos, A., Suarez, G. & Dominguez, L. (1999) *Int J Syst Bacteriol* **49**, 1263-73.
- 30 21. Van Soolingen, D., P.E.W. de Haas, J. Haagsma, T. Eger, P.W.M. Hermans, V. Ritacco, A. Alito, & J.D.A van Embden. (1994) *J. Clin. Microbiol.* **32**, 2425-33.
22. Samper, S., Martin, C., Pinedo, A., Rivero, A., Blazquez, J., Baquero, F., van Soolingen, D. & Van Embden, J. (1997) *Aids* **11**, 1237-42.
23. Mahairas, G. G., Sabo, P. J., Hickey, M. J., Singh, D. C. & Stover, C. K. (1996) *J Bacteriol* **178**, 1274-82.
- 35

24. Gordon, S. V., Eiglmeier, K., Garnier, T., Brosch, R., Parkhill, J., Barrell, B., Cole, S. T. & Hewinson, R. G. (2001) *Tuberculosis* **81**, 157-63.
25. Brosch, R., S. V. Gordon, K. Eiglmeier, T. Garnier, F. Tekaia, E. Yeramian, & S. T. Cole. (1999) in *Molecular genetics of mycobacteria*, eds. Hatful G. F. & Jacobs, W. R. Jr. (American Society for Microbiology, Washington, D.C.), pp. 19-36.
26. Radhakrishnan, I., K. M. Y., Kumar, R. A. & Mundayoor, S. (2001) *J Clin Microbiol* **39**, 1683.
27. Fletcher, H. A., Donoghue, H. D., Holton, J., Pap, I. & Spigelman, M. (2002) *Am. J. Phys. Anthropol.*, in press.
28. Mays, S., Taylor, G. M., Legge, A. J., Young, D. B. & Turner-Walker, G. (2001) *Am J Phys Anthropol* **114**, 298-311.
29. Nerlich, A. G., Haas, C. J., Zink, A., Szeimies, U. & Hagedorn, H. G. (1997) *Lancet* **350**, 1404.
30. Salo, W. L., Aufderheide, A. C., Buikstra, J. & Holcomb, T. A. (1994) *Proc Natl Acad Sci USA* **91**, 2091-4.
31. Rothschild, B. M., Martin, L. D., Lev, G., Bercovier, H., Bar-Gal, G. K., Greenblatt, C., Donoghue, H., Spigelman, M. & Brittain, D. (2001) *Clin Infect Dis* **33**, 305-11.
32. Parsons, L.M., Brosch, R., Cole, S. T., Somoskovi, A., Loder, A., Britzel, G., van Soolingen, D., Hale, Y., & Salfinger, M. (2001) in preparation

20

25

30

35